UNIVERSIDAD
COMPLUTENSE
MADRID

Proyecto de Innovación y Mejora de la Calidad Docente

Convocatoria 2015

Nº de proyecto: 49

Desarrollo de un Sistema de Recuperación de Información para el aprendizaje de buscadores

Responsable del proyecto: Juan-Antonio Martínez-Comeche

Facultad de Ciencias de la Documentación

Departamento de Biblioteconomía y Documentación

1.- **Objetivos propuestos en la presentación del proyecto.**

Los principales objetivos perseguidos en el proyecto se pueden resumir de la siguiente manera:

- Desarrollo de una herramienta informática que permita la asimilación de los fundamentos y principales técnicas de recuperación automatizada de la información por parte de un alumnado carente de una base tecnológica y matemática previa.
- Facilitar el aprendizaje autónomo de los factores y parámetros más relevantes relacionados con la recuperación de información.
- Favorecer la asimilación de objetivos docentes (en este caso, relacionados con la recuperación de información) mediante la experimentación, el análisis y la comparación de resultados obtenidos en casos particulares.

2.- **Objetivos alcanzados.**

Se han alcanzado la totalidad de los objetivos propuestos, resumidos de la siguiente manera:

- Se han seleccionado finalmente dos herramientas informáticas como elementos que permiten al alumnado (con escasa base tecnológica y matemática previa) la asimilación de los fundamentos y principales técnicas de recuperación automatizada de la información:
    - El programa MeTA (ModErn Text Analysis), desarrollado por el Departament of Computer Science de la University of Illinois con el objetivo expreso de facilitar la docencia e investigación en Recuperación de Información y en Minería textual, que incluye por defecto los modelos probabilístico y vectorial, y al que se ha incorporado el código correspondiente al modelo booleano.
    - El programa WauSearch, desarrollado por el profesor Manuel Blázquez Ochando, que incluye los tres modelos de recuperación de información.
- Se ha realizado ya un primer Taller de Recuperación de Información con los alumnos de tercer curso del Grado en Información y Documentación matriculados en la asignatura ñBúsqueda y Recuperación de Informaciónò durante el curso 2015-2016. Dicho taller ha incluido el manejo del programa MeTA para el aprendizaje de las principales técnicas de indexación (eliminación de palabras vacías, stemming o reducción morfológica y etiquetado de partes del discurso) y recuperación de información (modelo booleano clásico, modelo probabilístico clásico y modelo vectorial clásico).
- Simultáneamente a la realización del taller se ha procedido a la evaluación del mismo por parte de una alumna que ya había cursado la asignatura durante el curso 2014-2015, de manera que su opinión permita por una parte confirmar la mejora en la asimilación de objetivos docentes relacionados con la Recuperación de información (mediante comparación con la metodología seguida el curso anterior), y por otra parte la constatación de aspectos susceptibles de mejora en dicho programa.

3.- **Metodología empleada en el proyecto.**

En la elección y desarrollo de estas herramientas informáticas se pueden discernir las siguientes fases:

- REQUERIMIENTOS: esta primera fase comporta el establecimiento de las características que debe poseer la aplicación buscada. Dichas características esenciales ya han sido descritas previamente y se pueden resumir de la siguiente manera:
  - Posibilidad de elección del modelo de recuperación de información (booleano, probabilístico y vectorial)
  - Facilidad para la selección de los factores y parámetros empleados en la representación de los documentos de la colección, ya sean preexistentes o de creación por parte del alumno
  - Facilidad para la selección del algoritmo de equiparación consulta-documentos, ya sean preexistentes o de creación por parte del alumno

- ANÁLISIS: en esta fase se analizan los costes, el tiempo que ocuparía el desarrollo de las posibles soluciones y los beneficios que se obtendrían en cada caso.

- DISEÑO: en esta fase los dos profesores de la asignatura ñBúsqueda y Recuperación de informaciónòse ocuparán de diseñar el modo más eficaz de alcanzar la solución obtenida en la fase anterior, teniendo en cuenta esencialmente los siguientes aspectos:
  - Las características de los sistemas de recuperación de información de código libre ya existentes
  - El nivel de transformación de cada uno de estos sistemas para obtener la solución deseada
  - El tiempo y coste de implementación que conllevaría el empleo de cada uno de estos sistemas

- IMPLEMENTACIÓN: esta fase implica la contratación de una persona con formación en informática que llevará a cabo la codificación del diseño adoptado previamente a partir del sistema de recuperación de código libre más conveniente.

- PRUEBA: en esta fase final de evaluación del sistema en especial la alumna participante, a fin de comprobar en qué medida la herramienta cumple los requerimientos impuestos en la fase inicial.

4.- **Desarrollo de actividades.**

Las principales actividades desarrolladas a lo largo del proceso han sido las siguientes:

- Selección del programa MeTA por satisfacer plenamente los requisitos impuestos:
    - Programa de código libre.
    - Posibilidad de elección del modelo de recuperación de información (probabilístico y vectorial), con la salvedad del algoritmo correspondiente al modelo booleano, por otra parte el más sencillo de todos ellos. Para ello se contrató a un informático que llevó a cabo la codificación del modelo booleano en C++ y lo incorporó al sistema.
    - Facilidad para la selección de los factores y parámetros empleados en la representación de los documentos de la colección. Todos estos factores y parámetros se modifican en un único archivo de configuración del programa, con un simple editor de textos.
    - Facilidad para la selección del algoritmo de equiparación consulta-documentos. El algoritmo de equiparación se define en el mismo archivo de configuración del programa, junto con el resto de los factores y parámetros de análisis y evaluación.
    - Posibilidad de ampliación en el futuro a la docencia de las técnicas relativas a la minería textual.
- Desarrollo del programa WauSearch por parte del profesor Manuel Blázquez Ochando, con el ánimo de incorporar un buscador que utilizase bases de datos de tipo SQL, bajo distribuciones Apache, PHP y MySQL, especialmente apto para las búsquedas en Internet. Ello ha dado lugar finalmente, como herramienta de enseñanza en las técnicas de búsqueda de información, al desarrollo de un buscador global de la Web que proporciona un entorno de aprendizaje real, con herramientas y operadores de filtrado avanzados. Algunos de los aspectos más destacados del programa WauSearch son los siguientes:
    - Programa de código libre.
    - Incluye los más importantes procesos de análisis y representación de documentos, destacando la normalización de los textos, la indexación, la supresión de código fuente, la tokenización, la transliteración de caracteres y la eliminación de palabras vacías.
    - Posibilidad de empleo de simuladores de modelos booleanos, vectoriales y probabilísticos.
    - Facilidad para la evaluación de algoritmos de recuperación.
    - Posibilidad de empleo de programas parser de recuperación de metadatos.
- Realización de un primer **Taller de Recuperación de Información** con los alumnos de tercer curso del Grado en Información y Documentación matriculados en la asignatura ñBúsqueda y Recuperación de Informaciónò durante el curso 2015-2016. Dicho taller ha incluido el manejo del programa MeTA para el aprendizaje de las principales técnicas de análisis e indexación (eliminación de palabras vacías, stemming o reducción morfológica y

etiquetado de partes del discurso) y recuperación de información (modelo booleano clásico, modelo probabilístico clásico y modelo vectorial clásico).

- Evaluación del programa MeTA por parte de una alumna que ya había cursado la asignatura durante el curso 2014-2015, de manera que su opinión permita por una parte confirmar la mejora en la asimilación de objetivos docentes relacionados con la Recuperación de información (mediante comparación con la metodología seguida el curso anterior), y por otra parte la constatación de aspectos susceptibles de mejora en dicho programa. El resumen final de dicha evaluación realizada por María Inmaculada Aguirre Artigas es el siguiente:
  ñEl objetivo del Taller ha sido el análisis de los aspectos principales de las búsquedas realizadas en distintos modelos clásicos de recuperación de información: booleano, probabilístico y vectorial. El ejercicio se dirige a alumnos matriculados en la asignatura de Sistemas de Recuperación de Información de 3º de Grado.
  La metodología docente de la práctica se plantea mediante la presentación del profesor en el aula del sistema de recuperación MeTA que permite emplear los tres modelos clásicos de recuperación de información para así poder comprobar su funcionamiento y realizar comparaciones entre ellos.
  Los conocimientos previos de los estudiantes son conocimientos teóricos impartidos en el aula que les habrán proporcionado la base que les permitirá realizar esta práctica además de un primer taller en el que ya han trabajado con el sistema MeTA (instalación y factores de análisis y representación de documentos); quizá el arranque del programa sea la parte que puede resultar de más difícil comprensión para aquellos estudiantes que no estén familiarizados con la utilización de máquinas virtuales de simulación de sistemas operativos.
  Todo el ejercicio se realiza bajo la tutorización del profesor en el aula y con el apoyo de un documento en el que se detallan los pasos a seguir; además el propio profesor acompaña a los alumnos en su desarrollo guiándolos mediante la proyección en pantalla del ejercicio práctico que él mismo lleva a cabo en su ordenador.
  El taller es innovador en el sentido de que pretende ir más allá de la mera transmisión de conocimientos teóricos por parte del docente y aunque, si bien es cierto, esa transmisión previa de conocimiento es imprescindible; la práctica permite al alumno observar en primera persona cómo funcionan los distintos buscadores tradicionales mediante la realización de consultas idénticas en una colección de documentos compuesta por 23.566 documentos (moocs de temática de carácter académico muy diversa) y 192.269 términos únicos.
  La metodología fomenta la participación activa del alumno y propone la observación y reflexión individual posterior mediante la realización de un ejercicio en el que se deberá ñlanzaròal sistema cuatro búsquedas distintas y anotar cuáles son los tres documentos más relevantes que el buscador devuelve, el tiempo que invierte en la realización de la búsqueda y la similaridad del documento recuperado con la consulta realizada.
  Los distintos pasos que el alumno va dando en la configuración del buscador de prueba permiten observar en la práctica cómo se procede a la creación de un fichero inverso y qué decisiones hay que tomar para ello: colección sobre la que va a trabajar; procesos previos que se efectuarán sobre los documentos

(eliminación de palabras vacías y stemming) y mediante qué técnicas. La observación en pantalla del proceso de indexación nos informa sobre ésos procesos aportando datos concretos sobre la información contenida en el fichero: número de documentos, longitud de los mismos, cantidad de términos únicos indexados y tiempo necesario para la creación del fichero.
En este taller se realizó la emulación de dos de los modelos clásicos de recuperación de información: el probabilístico y el vectorial, debido a problemas de última hora con el fichero que incluía el modelo booleano añadido al programa MeTA; se anotaron los tres primeros documentos que en cada uno de los sistemas se devolvieron como más relevantes, así como el tiempo que invirtió el sistema en su recuperación (eficiencia) y su grado de similaridad con respecto a la consulta realizada por el usuario.
No cabe duda de que la práctica muestra al alumno el funcionamiento de los sistemas de recuperación de información; la información que el sistema devuelve y la que el propio alumno configura: elección de proceso de reducción para la creación del fichero inverso; tiempo invertido en la búsqueda y recuperación; documentos relevantes ordenados por grado de similaridad y los propios documentos recuperados son un ejercicio necesario para facilitar la comprensión de los conocimientos teóricos impartidos en el aula.ò